Development Considerations for Implementing a Voice-Controlled Spacecraft System

George Salazar, *NASA Johnson Space Center, Houston, Texas 77058*

Abstract— **As computational power and speech recognition algorithms improve, the consumer market will see better-performing speech recognition applications. The cell phone and Internet-related service industry have further enhanced speech recognition applications using artificial intelligence and statistical data-mining techniques. These improvements to speech recognition technology (SRT) may one day help astronauts on future deep space human missions that require control of complex spacecraft systems or spacesuit applications by voice. Though SRT and more advanced speech recognition techniques show promise, use of this technology for a space application such as vehicle/habitat/spacesuit requires careful considerations. This paper provides considerations and guidance for the use of SRT in voice-controlled spacecraft systems (VCSS) applications for space missions, specifically in command-and-control (C2) applications where the commanding is user-initiated. First, current SRT limitations as known at the time of this report are given. Then, highlights of SRT used in the space program provide the reader with a history of some of the human spaceflight applications and research. Next, an overview of the speech production process and the intrinsic variations of speech are provided. Finally, general guidance and considerations are given for the development of a VCSS using a human-centered design approach for space applications that includes vocabulary selection and performance testing, as well as VCSS considerations for C2 dialogue management design, feedback, error handling, and evaluation/usability testing.**

I. INTRODUCTION

Since the first commercial use of speech recognition technology (SRT) in the 1980s—primarily in the telephone industry—SRT has improved concerning speaker recognition performance and dialogue design. Today, many commercial industries, including banking, airline reservations, medical, automotive, and legal (to name a few) use SRT. The iPhone industry has advanced human-computer interface (HCI) dialogue development using artificial intelligence (AI) and natural language processing (NLP). It will soon become people's chief means of interacting with computing devices and systems. For future space missions, these advancements may one day help astronauts control complex spacecraft systems, information storage, and retrieval, or spacesuit controls by voice [6].

Currently, more than 50 people in the Mission Control Center (MCC) support the operation of the Space Station. They serve as an extra pair of eyes to the astronaut overseeing the numerous station systems and handling emergencies. As humanity advances farther from Earth, the round-trip communication delays between Earth and the spacecraft increases, requiring more onboard MCC functionality to maintain and control the vehicle. Hence, deep space missions to Mars and beyond will place challenging constraints on the crew that will demand an efficient and effective HCI to control a highly complex vehicle/habitat system. NASA Human Research Program (HRP) has identified inadequate HCI as a risk for future missions. New tools and procedures to aid the crew in operating a complex spacecraft will be required. Voice-controlled spacecraft systems (VCSS) could potentially help solve the HCI risk.

Though SRT has come a long way, it still struggles with challenges such as background noise and human speech variability influenced by task loading and the user's psychological and physiological health. SRT recognizes the words spoken but lacks an understanding of the meaning of the words. The responsibility for interpreting the users meaning lies with the VCSS application dialogue between the user and the machine. However, the machine-understanding dialogue is still a long way from understanding the speaker's intended meaning. Incorrectly recognized words are treated cautiously as the system could branch into a wrong part of the application software code-- furthering the likelihood of more recognition errors. Implemented correctly, VCSS in a data entry application permits increased machine or system operator efficiency with a high level of recovering from recognition errors (RE). If incorrectly developed, the task workload can increase, resulting in potentially more REs, making the system unacceptable to the user.

There are numerous published reports, masters'/ doctoral thesis, and presentations on the development and implementation of SRT applications. The majority of these documents focus on commercial, aviation, or automotive applications rather than spaceflight. The few industry standards in the military [16] and the automotive industry provide some design guidelines that may be applicable for space applications. However, the space environment is different from any of these industries. NASA has provided some guidance on applying speech recognition [17] but lacks in details. Design and use of this technology for space applications, such as vehicle/habitat/spacesuit, require careful considerations, particularly when it comes to user satisfaction and safety.

This paper provides considerations and guidance in developing a VCSS for space applications. Specifically, it addresses C2 dialogue applications associated with controlling a subsystem by voice (such as a camera or a robot system), dictation, or retrieving information such as system status or data files. In C2 applications, the user initiates the dialogue (user-initiated) rather than the machine (computer-initiated). In this paper, spacecraft system can be an application in a habitat, (in orbit or on the surface), spacesuit, or the spacecraft vehicle carry the crew to its destination.

First, current SRT limitations as known at the time of this report are given. Then, highlights of SRT used in the space program provide the reader with a history of some of the applications used in space research. Next, a short overview of the speech production process and the intrinsic variations of speech are provided. Then, general VCSS development guidance using a human-centered design approach is given. Finally, considerations for vocabulary selection, C2 dialogue, feedback, error handling, and evaluation/usability testing are discussed.

II. SPEECH RECOGNITION PROCESSING LIMITATIONS IN SPACE

Several challenges exist for applying SRT to deep space human missions. High-performance computing platforms such as used for terrestrial application are not suitable for deep space applications. Unlike low-earth orbit where the Earth's magnetic field provides important protection of a spacecraft and its electronics from damaging radiation, outside the magnetic field, the radiation is much more severe [9]. The galactic radiation can cause electronics (if not designed for the deep-space environment) to experience single event effects such as functional interrupts resulting in an abrupt halt to operations and possibly catastrophic failures.

Environmental noise also plays a significant role in speech recognition performance, specifically how well the system recognizes speech in the presence of stationary and non-stationary noise. Current technology does well when the signal-to-noise ratio (SNR) is better than 25 dB, such as in an office environment. However, low SNR causes poor recognition results [2]. The combination of enormous reverb such as in a spacesuit and both stationary and non-stationary noise (such as pumps kicking on and off) poses challenges for SRT in space applications. Use of deep neural networks to predict the acoustic/noise environment has shown promising results [8]. However, commercially-available General Purpose Graphics Processing Unit (GPGPU) are used which are not suitable for the deep space environment.

Natural language processing (NLP) is still maturing. It still struggles with correctly parsing a sentence and understand its meaning. Therefore, it will not be considered in this paper. (Note: The amount of hardware and software necessary for NLP would place a significant burden on a spacecraft's size, weight, and power constraints. For example, Apple's Siri uses the iCloud and several servers loaded with GPGPUs for its speech recognition function.). Even in human-to-human communications, there are misunderstandings in what was said. Factors such as prosody of the speech indicate grammatical structures and stress of a word that suggests importance in the sentence. The ambiguity of the English language is also another challenge. Homophones (words that sound the same but have a different meaning) can confuse the NLP application. Similarly, word boundary ambiguity recognition can pose speech recognition difficulties (e.g., recognize speech vs. wreak a nice beach) [1].

III. SPEECH CONTROL DEMONSTRATIONS AND FLIGHT EXPERIMENTS

NASA has been involved in SRT for space applications for quite some time for both spacecraft and ground operations. A few SRT investigations have been performed in a spacecraft operational environment. All voice-controlled investigations have been in the form of C2 applications. What follows is a summary of select SRT investigations and demonstrations that have occurred in the space program-both ground and space.

A Feasibility of Controlling the Space Shuttle Video System by Voice

In the late 70s, the Jet Propulsion Laboratory (JPL) developed a ground demonstration of a prototype voice controller to manipulate the Space Shuttle Closed Circuit Television (CCTV) System. It provided an opportunity to investigate the potential of the technology to improve the performance of the Space Shuttle remote manipulator system (RMS) operations [10]. Several successful ground tests occurred at the JSC Space Shuttle Manipulator Development Facility (Figure 1) to

perform berthing, maneuvering, and deploying a simulated science payload in the Shuttle payload bay. The conclusion was that a voice-controlled system would help reduce space shuttle robotic operations workload. [4]. This ground test led to the eventual flight demonstration that will be discussed later in the paper.

Figure 1. Manipulator Development Facility

B. EVA Voice-Activated Information System

In 1990, a ground demonstration of a voice-activated information system was developed to control a heads-up display on a prototype EVA suit. Figure 2 shows the prototype system. The system was designed to assess the effectiveness of a voice-activated heads-up display as compared to a manual/electronic wrist-worn or "cuff" checklist during EVA operations. The issue with a manual cuff checklist is that the wrist is not always in the best position to read a page. Also, at least one hand is required to turn the pages. Electronic cuff-mounted checklists have issues with size and displaying data in the presence of sunlight. The voice system was termed an improvement over the manual/electronic cuff checklist allowing both hands on the job while moving through procedures by voice. However, there was an issue with disruptions between the voice communication with the ground and other astronauts versus communication with the SRT [5].

Figure 2. Prototype Heads-up display

C. Front-end Processing of Speech in a Noisy Environment

In 2005, JSC human factors with the help of avionics engineering jointly worked on a project to investigate ways to reduce noise from spoken speech in a spacecraft/spacesuit. The noise in the spacecraft can be as high as 80-85 decibels sound pressure level (dB SPL) that includes fans, mission control chatter along with impulse noise from pumps and circuit breakers kicking off and on. In addition, the reverberation can be challenging as well. A speech enhancement system using Independent Component Analysis to separate the speech from noise and background chatter was used as the front-end audio processing. The algorithm ran on a digital signal processor that was in line with the dual microphone signals from the headset and to the laptop that housed the speaker-independent recognition software. Figure 3 shows the system components. The results were mixed. For males in an 80 dB SPL background noise, the speech enhancement enabled should a significant

improvement over the enhancement not enabled. For females, there was a decline in performance that may have been attributed to the speech recognition engine tuned for males and not for females.

Figure 3. Front-end Enhancement Processing System

D. STS-41: Voice Command System Flight Experiment

The results of the JPL-developed voice control of the Shuttle CCTV system provided an opportunity for the Johnson Space Center to pursue development of a Voice Command System (VCS) flight experiment demonstration to control the Shuttle CCTV system—the first-ever control of a spacecraft system by voice [25]. The objective of the flight experiment on STS-41 was to collect baseline data on the effect of microgravity on speech production and recognition, as well as show the utility and operational effectiveness of controlling a spacecraft system by voice.

The system utilized a military speaker-dependent commercial recognizer developed by Spacecraft Incorporated out of Huntsville, Alabama. The unit included a commercial speaker-dependent recognizer made by Votan. The unit required capturing and storing into non-volatile memory command word templates used in the speech recognition process. At power-up, the astronauts had to identify themselves by speaking their name. The unit would them retrieve their templates from non-volatile memory. Figure 5 shows the elements of the system and the block diagram.
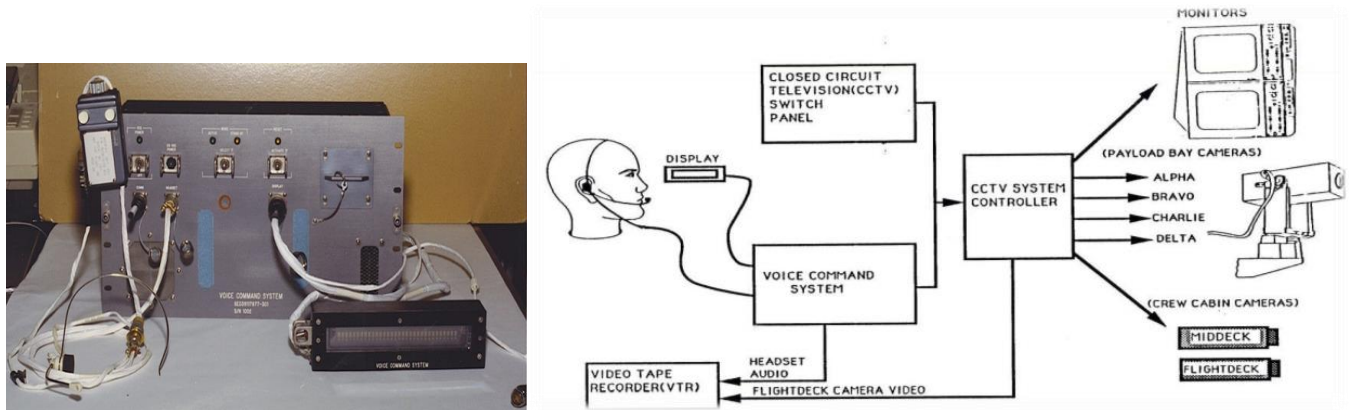

Figure 4. Voice Command System and Block Diagram

The VCS paralleled the Shuttle CCTV switch panel that was comprised of 33 switches, both toggle, and push-button switches to control the cameras, monitors, and pan/tilt units. Twenty-three command words mimicked CCTV switch panel closures. Two command words, "stow cameras," was used as a macro command that would configure all four payload cameras into their stow position, automatically. The front panel comprised of the power switch, circuit breaker, toggle to put the VCS into active or standby mode, reset switch, Shuttle headset, display and audio output to the recorder and the storage compartment for the two astronaut template EPROMs.

The recognition accuracy of the command word templates captured on the ground dropped about 10% for one astronaut and about 30% for the other. Several factors played into the lower performance, including microphone quality (used existing Shuttle headset) and the operational environment. However, both astronauts were highly successful in creating on-orbit voice templates and obtaining a recognition accuracy average of 97%. Both astronauts agreed that the system would be attractive for use during CCTV mission operations, especially macro commands.

E. Detailed Test Objective 675 "Voice Control of the Shuttle CCTV System

A follow-on flight of the VCS was approved for STS-78. STS-78 was a 16-day mission compared to the 5-day STS-41 mission that provided an opportunity to examine if a longer duration mission affects the speech process. Two astronauts used the system early, midway through and toward the end of the mission. Several modifications were made to the electronics in response to requests by the astronaut office to improve the VCS flown on STS-41. Most notably, the updated system was a commercially available speaker-independent engine with adaptive speech recognition technology that was cutting-edge at the time. The recognizer outputted the top three words in the vocabulary it thinks was spoken along with confidence scores of each. For questionably recognized words, a query in the dialogue would invoke a "yes" and "no" response to determining what word was said by the astronaut to update the word template. The adaption feature would track changes to the astronaut's voice. Other modifications included adding a microphone gain feature to help compensate for microphone placement, and four macro commands were incorporated in addition to the "stow cameras" macro from STS-41.

On additional feature added to the system was in-flight raw audio recordings of the two astronauts that used the system. These recordings were stored for ground analysis to determine if microgravity affected the speech process. Preflight audio samples were obtained from both astronauts. Post-flight speech analysis showed that one of the astronaut's fundamental frequency changed much more than the other astronaut. In fact, on the 16th mission day, the astronaut had difficulty with speech recognition that may have been attributed to this frequency shift even though the pitch period remained steady.

Overall, the adaption feature worked reasonably well, providing average recognition accuracy for both astronauts of 90% throughout the mission. However, the VCS had limited capabilities in adjusting for placement of the microphone. One identified shortcoming was that the macros did not provide feedback to the astronauts, such as percent complete, nor did they provide the ability to stop/resume the macro. The cabin acoustics environment provided unique challenges as 95 dB SPL audio messages were sometimes uplinked, and the reverb in the cockpit confused the VCS. During pan/tilt commanding, the querying feature for questionable recognition interrupted the task and therefore is not recommended. Another astronaut observation was that speech recognition for continuous control, such as tracking a location on Earth with cameras, is difficult to do compared to the traditional method of using the switch panel [26]. As experience has shown, crew members will revert to the primary control system to which they are used to if an alternative method is not sufficiently accurate or usable [13].

1. Flight Experiment Command Word Selection and Vocabulary Diagram

For both STS-41 and STS-78, the command words were nearly the same except for Confidence Check, Query Check, and User Setup. Macros had their own node. The STS-78 crew modified some of the CCTV command words to make it easier for them to remember. Not all CCTV switch panel buttons were implemented, just the most commonly used ones. Macro commands were considered useful by both crews, which permitted one command word to cause a plurality of commands sent to the cameras. The table 1 below shows the command words used that emulated the CCTV switch panel and unique commands, and Figure 5 is the nodal vocabulary structure. Note in Figure 5 that the transition words "voice command" was spoken first and then within 2 seconds "activate" to put the system in the listening mode. If they did not speak thw command word "activate" within 2 seconds, a timer would time out, and the system would return to first listening for "voice command."

Table 1. Voice Command System Vocabulary

| A7 Panel Commands	Purpose	VCS Unique Commands	Purpose
Up/Tilt_Up	Tilts camera up	Confidence_Check	Trains all words
Down/Tilt_down	Tilts camera down	Query_Check	Trains "Yes"/"No"
Zoom_In/Close_Up	Zooms camera in	User_Setup	Loads requested parameters
Left/Pan_Left	Pans camera left	Setup	Transition word
Right/Pan_Right	Pans camera right	Voice Command	Transition word
Zoom_out	Zooms camera out	Activate	Transition word
Focus_near	Focus camera near	Stop/Halt/Whoa	Stops camera/pan/tilt movements
Focus_Far	Focus camera far	Macro	Transition word
Monitor_1	Selects monitor 1	Configure	Transition word
Monitor_2	Select monitor 2	Standby/Relax/ Go_To_Sleep	Put VCS into non-listening
Alpha	Select camera A	Move/Adjust	Transition word
Bravo	Select camera B	Macro_1	Reserved
Charlie	Select camera C	Macro_2	Downlinks test pattern
Delta	Select camera D	Macro_3	Views earth using cameras A,D
Middeck	Select Middeck camera	Macro_4	Scans PLB port side w/cameras A/D
FlightDeck	Select Flightdeck camera	Macro_5	Scans PLB STBD side w/cameras C,D
1_Left	Select Monitor1left side	Calibrate	All 4 PLB cameras placed into known position
1_Right	Select Monitor1right side	Stow_Cameras	All 4 PLB cameras into ascent/reentry position
2_Left	Select Mon 2 left side	Yes	Query response for choosing correct word spoken
2_Right	Select Mon 2 right side	No	Query response for not choosing a word spoken
Open_Iris	Opens the camera iris	1-,2-, 3-,5-,10,15- degrees	Number of degrees selected for a camera lens or
Close_Iris	Closes the camera iris		pan/tilt movement command
Change_Rate	Changes the pan/tilt rate		
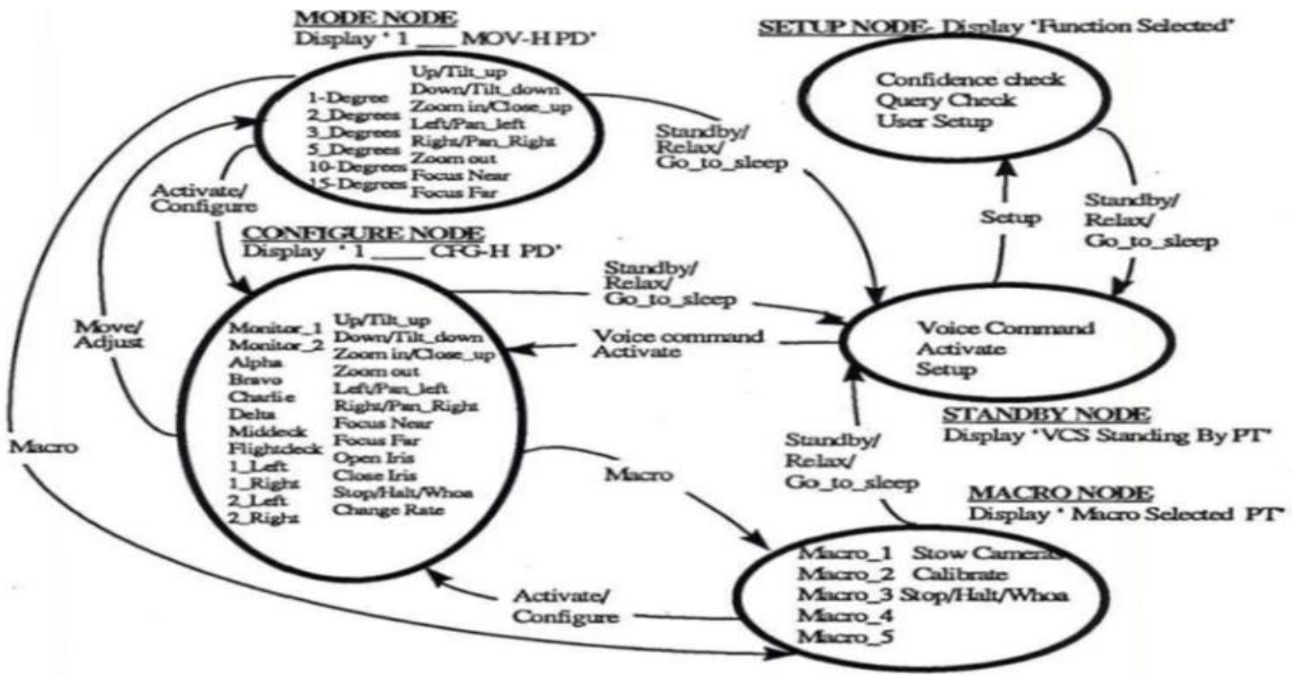
Figure 5. Nodal Vocabulary Structure

IV. SPEECH VARIATIONS

The human speech process contains variations associated with fundamental physiological and/or psychological factors. These differences can be the result of intrinsic human variations, such as vocal tract cavity changes, and/or changes due to the environment, physical, and cognitive stresses. A short introduction to this topic follows.

A. Human Intrinsic Speech Variations

Speech recognition or a natural language understanding system will never achieve 100% recognition and understanding; Even humans cannot attain that. In fact, a speech recognition system is similar to a person with hearing problems such that hearing worsens with higher noise levels. The purpose of SRT is to detect and categorize a sound wave coming from a microphone into one of the basic units of speech, namely word comprised of phonemes. In the English language, there are about 40 phonemes while other languages have more or less. The challenge for the speech recognition system is to detect these phonemes, compare them against other similar phonemes using statistical models, and match them against a known library comprised of known words, phrases, and sentences.

Figure 6 shows the speech production vocal organs associated with the vocal tract. The lungs, diaphragm, and stomach serve as the air stream source through the larynx. The larynx includes the vocal cords that modulate the airflow from the lungs. This modulation creates a quasi-periodic pressure wave impulse with a specific frequency determined by muscle tension force and the mass and length of the vocal tract. These factors vary by age and gender. The vocal tract cavities filter the air stream between the glottis and the lips. The lips, tongue, and teeth further modify the air stream. The result is a sound wave emanating from the mouth as either voiced or unvoiced speech signal.
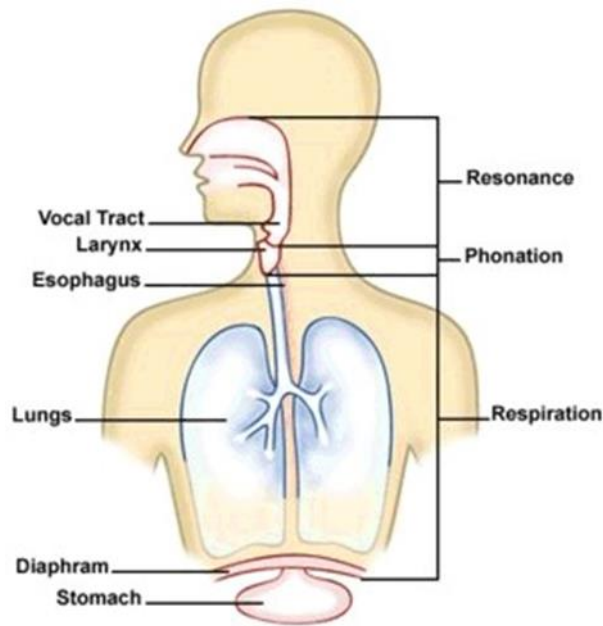
Figure 6 Voice Production (Credit: Johns Hopkins Voice Center)

Unvoiced speech is aperiodic with low energy compared to voiced speech. For voiced speech, both the vocal chords and the vocal tract are used. Consequently, voiced speech is periodic and is most important for intelligibility. The energy of voiced regions is an order of magnitude larger than of unvoiced regions [2]. An example of voiced speech would include vowel sounds. Example of unvoiced would be like the consonants /p/, /s/, /t/, or /k/. Also, some nasal consonants such as /n/ and /m/ are not as energetic as voiced sounds. The low energy can pose a problem with SRT trying to determine the beginning of a word/phrase. The problem is further exasperated with soft-spoken speakers. For applications that use far-field microphones such as array microphones, background noise can worsen the issue, especially for unvoiced speech. Physiological and psychological factors, as well as fatigue from speaking too long, can further alter a person's speech intrinsic variations.

B. Effects of the Space Environment on Speech

The spacecraft environment and the mission tasks present challenging problems for speech recognition. In addition to intrinsic speaker variations, the environment, physical, and cognitive stress in a space mission can further affect the speech process and recognition. For example, the microgravity environment causes no sinus drain and micro-gravity causes muscular atrophy that could alter the speech process.

To date, only a few experiments/investigations regarding the speech process in the space environment have occurred [7, 25, 26]. Spacecraft volume and layout affects the reverb of a speech signal, which can cause smearing of the spoken words [14]. Spacecraft systems typically have loud uplink and/or caution and warning audio message capabilities that can leak into the microphone and corrupt speech recognition. A loud cabin environment can create a Lombard effect, causing the crewmember to speak loudly over the background chatter/noise [11]. EVA activities can put physical stress on the crewmember that can alter their breathing/speech. The same can occur with tasks that have a high cognitive workload [12]. In short, the many human intrinsic speech variations pose challenges for spacecraft applications using speech control and natural language processing. The next section provides design considerations to help mitigate these recognition challenges.

V. VOICE-CONTROLLED SPACECRAFT SYSTEMS (VCSS) DEVELOPMENT CONSIDERATIONS

Developing a VCSS involves several human factors engineering (HFE) fields, including physiology, psychology, and linguistics. It also comprises computer science, electronics and acoustics engineering, and signal processing. Human-in-the-loop evaluation is also necessary to optimize/refine the system. Since astronauts are highly skilled with knowledge of the task through training, this report focuses on considerations for implementing a C2 interactive speech control system for space applications. Following a description of contributing factors that can cause the VCSS to fail to complete a task/goals successfully and the Human-Centered Design (HCD) process recommended for the development of a VCSS application, VCSS design guidance in vocabulary development, handling errors, dialogue design, and usability evaluation and testing are given.

A. Factors Affecting the VCSS Task/Goal

Implementing a speech recognition application for the space environment can be challenging, particularly in a demanding environment such as a spacesuit where reverb, suit noise, and external communications and suit alarms are present. Even in a

habitat, rover, or spacecraft, the noise and reverb effects alone make good speech recognition difficult. Coupled with the user physiological and psychological state of the user, and it can be a very challenging situation for the VCSS. Figure 7 shows a fishbone representation of the factors that can affect the VCSS to successfully completing a portion or all of a task. Complexity/workload is not just how complex the task is that affects workload but also the added cognitive workload that can happen by the use of voice control (Note: Speech shares short term memory and thus affects cognitive problem-solving and memory recall. [27]). Redundancy in the task is the backup system in case the recognition system is not working properly. In the case of the Voice Command System (VCS) on Shuttle, the CCTV switch panel was the backup to the VCS. Habitability is associated with the place the task takes place such as the spacecraft, the rover, or the spacesuit. Note that the habitability acoustic environment is not just the reverb that can affect speech recognition but also the habitat temperature and how crowdedness can add to the stress of the task.
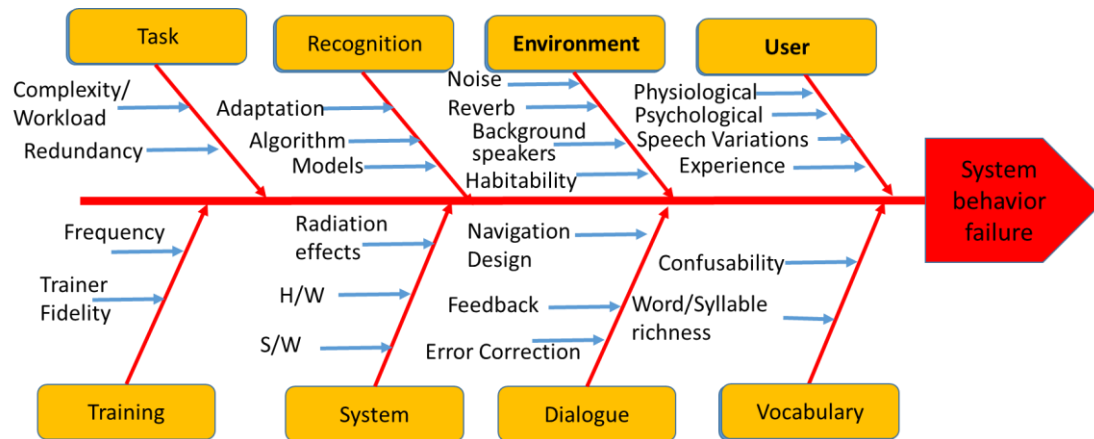

Figure 7. Factors Affecting System to Complete a Task Successfully

Training frequency can affect performance as well. If the crew has not trained sufficiently, there is a possibility that the crew can forget a step/spoken command that takes the application to another undesired state. If the system does not provide correct feedback, the user could be trying to command the system with words not in the vocabulary for that state. Caution must be exercised with systems using adaptation as it can skew to a certain individual's voice that affects the other user's speech recognition performance. The models may also affect the performance if the acoustic model not tuned for the environment. For a front-end speech enhancement hardware implementation such as a mic-array, if the front-end cannot handle impulse noise or reverb effects, the corrupt speech signal will propagate to the recognizer resulting in a misrecognition/error. A key aspect of the design is the dialogue features such as feedback and error correction. More will be said later in the paper about this.

B. Human-Centered Design (HCD) and the Development Process

Speech recognition applications can be quite human-machine interaction-intensive, especially in a multi-modal application where both voice and another modality, such as gestures, is used. Developed incorrectly without factoring the human element can result in an unusable system. Therefore, as part of the systems engineering effort, an HCD process is the recommended development approach. The HCD approach ensures the VCSS meets requirements and crew acceptance of the system. Also, it is highly recommended that a human systems integration (HSI) practitioner participates in the development early in the project lifecycle to ensure the human element is factored into the design of the system. Critical aspects of the human-centered effort are [18]:

- The design is based upon a clear understanding of the crew, the task, and the environments.
- The crew is involved throughout the design and development process.
- The design and requirements are refined by evaluation of the prototype system.
- The process is iterative.

For flight systems, the NASA systems engineering process per NPR 7123 is used. Figure 8 shows an abbreviated NASA systems engineering project life-cycle phases along with key project milestone reviews. Most of the HCD occurs in the formulation phases leading up to critical design review. During the implementation phase, design evaluation becomes formal usability testing with the flight design as the system components are integrated with the rest of the system and requirements verified. Finally, system requirements validation to certify the system for flight is performed. Throughout all the phases, the HSI practitioner ensures all aspects of the system and human in the given environment are optimized.

Figure 8. NASA Life-Cycle Phases with Project Milestones

HCD activities are part of the systems engineering process. Depending on if the VCSS is a research/development project or a flight system, the process may differ. A key element of the HCD process is user evaluations that feedback into the development effort. Figure 9 shows the iterative HCD activities adopted from ISO standard 9241-210 [18]. Most of the HCD effort should occur prior to critical design review (CDR) where design change costs are minimal.

The figure does not suggest a linear process but rather illustrates how each activity uses outputs from other activities. A short description of each activity follows.
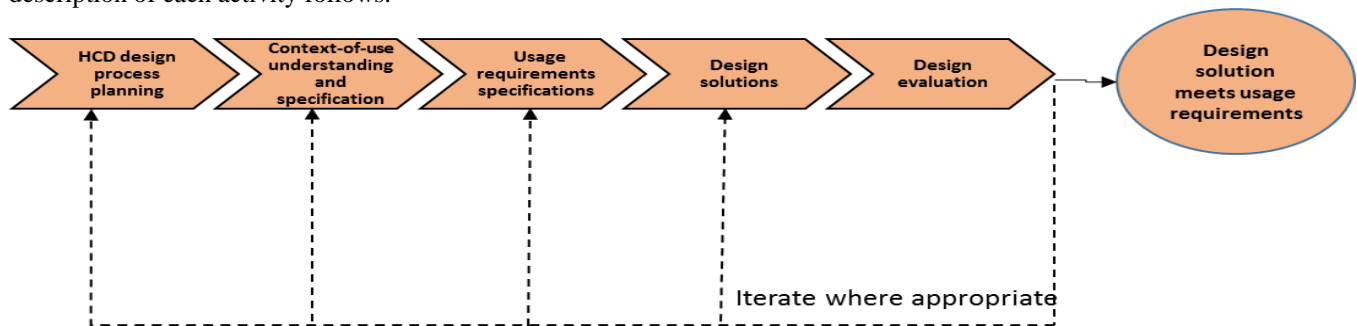

Figure 9 Human-Centered Design Process

1. HCD design process planning

This activity is ideally led by an HSI practitioner with the main objective of developing an HSI plan that integrates HCD into all phases of the VCSS project, as well as ensuring all HSI domains are addressed. The process includes requirements development, project resources, risk assessment, and scheduling of HCD activities that factor adequate time for evaluations and design iterations/optimization.

2. Context-of-use

The key activities include working with stakeholders to define the goals and tasks of the systems. HFE personnel are heavily involved in cognitive engineering assessment and analysis of the task, the crew, and system all working together. A concept of operations (ConOps) is used to define nominal and contingency use cases and task scenarios of what the user must do regarding actions and/or cognitive processes in the given environment to achieve the overall task. The use cases/scenarios help refine VCSS requirements as well as build the preliminary task vocabulary. They also help describe the task functions that need to be performed, which defines the human-machine interaction behavior. As evaluations of the model (whether it is a wireframe, prototype or mockup) continue, feedback from the crew provides updates to the ConOps, scenarios, use-cases, VCSS requirements, and design.

3. Usage Requirements & Specifications

Based on the task and functional analysis, usage requirements are developed along with the system requirements. Then decomposition of system requirements is performed to further define the VCSS. Many of the functional/sub-functional requirements are stated regarding performance such as time to perform a sub-task or speech recognition processing time (spoken utterance recognized to system output). Based on the functional requirements, a functional architecture is developed showing the functions and sub-functions. All requirements should be allocated to the functions/sub-functions of the VCSS.

Also, requirements are allocated to either the hardware, software, or the user. In the design phase, the design is compared against the system requirements to assure compliance or to re-evaluate the design/requirements.

4. Design Solutions

This activity focuses on creating a design that meets all requirements, both the functional and non-functional requirements (such as weight, power, and size). There are three main development activities involved to develop a VCSS, namely VCSS design, VCSS dialogue design, and Human Factors Engineering (HFE). All three are performed interactively. Figure 4 shows a Venn diagram of the three main development activities interacting to create and optimize the VCSS application. The acoustic environment is factored into all VCSS design considerations. Note that HFE is heavily involved in the design evaluation as well as the design solutions.

5. Design Evaluation

Design evaluation is a key component of human-centered design that focuses on a design that meets the task goals and objectives while ensuring the system complies with users' capabilities. The evaluation provides user performance measures and subjective feedback (both qualitative and quantitative) that can be used to improve the system throughout the engineering design life cycle. The evaluation offers designers and engineers with direct feedback from the earliest stages of design through system formal usability verification and validation testing and operational use. The preference is to perform several design evaluations with high-fidelity prototypes when the cost of design changes are minimal.

6. VCSS Design Solutions and Evaluation Interaction

Figure 11 shows a Venn diagram of the three main development activities interacting within the HCD Design Solutions and Evaluation activities to create and optimize a VCSS application. The acoustic environment is factored into all VCSS design considerations.
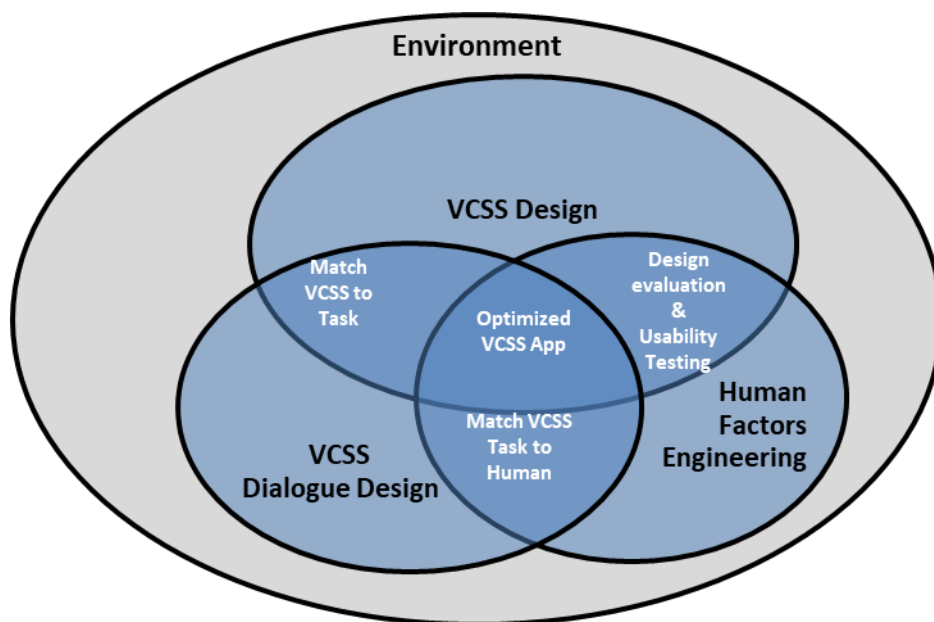

Figure 11 VCSS Design & Development Process Venn Diagram

The VCSS design effort develops the hardware and software/firmware necessary to support the task in the given environment. Optimization of the VCSS hardware/software to the task occurs between the VCSS design and the VCSS task design activities such as determining if a backup control is needed or type of feedback. Also, the speech recognition engine that will be used is selected (see confusability matrix testing). Based on the task use cases/scenarios, the C2 dialogue algorithm (see C2 Dialogue Design) is developed through the VCSS dialogue design and VCSS design efforts. As part of the VCSS design effort, the hardware computing platform is matched to the speech recognition engine selected for the task. The type of microphone (head-worn or array) and the front-end noise/echo suppression algorithm are developed as well. HFE works with the VCSS design engineers to develop and perform numerous evaluations and tests with the crew to optimize the VCSS design with the human regarding usability and crew acceptability. This optimization may include minimal workload, response time, and intuitive C2 dialogue navigation. HFE also supports the development of the VCSS task and vocabulary selection to ensure an optimized VCSS experience for the user. Note that, depending on the phase of the project, there can be

different levels of effort in these development activities. Also, these activities feedback to HCD planning, context-of-use, or user requirements and specifications-particularity, up to CDR.

VI. KEY VCSS GUIDANCE ACTIVITIES

Given that a good speech recognition engine(SRE) has been selected, and the hardware has the throughput for the SRE and front-end noise reduction performance, the key to developing a successful VCSS is focusing on the vocabulary for the task, the dialogue design that navigates the user, and the error correction scheme to handle questionable recognitions gracefully. The dialogue design provides the "look-and-feel" of the interface that can provide either acceptance or rejection of the system by the user. What follows are guidelines and recommendations for developing the vocabulary, dialogue design, and the error correction scheme.

A. Vocabulary

Task vocabulary selection is both an art and a science. Words that sound similar to a human may be recognized differently by SRT. SRT statistical recognition algorithms can produce artifacts in the recognition process that may require replacing words that otherwise appear would be good candidates. (Note: It is highly recommended to engage the speech recognition vendor in the vocabulary selection process, particularly reporting issues encountered during the performance evaluation of the vocabulary. They have the tools to optimize recognition parameters.) The following are considerations in selecting the vocabulary.

1. Task-Relevant

Vocabulary selection is based on the functional requirements levied on the system and the system ConOps that spells out the end goal(s). From these two documents, use cases or scenarios drive task development that must meet the requirements and end goals. The vocabulary should be natural and reflect the task actions the system must do, e.g., command a camera pan/tilt unit should have the vocabulary words "pan" and "tilt" with the associated "left" "right," and "up/down" respectively. It should also be distinctive to minimized recognition errors. However, performance testing may determine that an alternate task vocabulary word(s) may be required.

2. Phonetic Features

During the selection of task-related vocabulary words, the design team should be aware of vocabulary phonetic features that could drive the use of specific vocabulary words [3]. Most words contain voiced and unvoiced consonants (or consonant pairs) that may determine the use of certain words. Voiced consonants such as /g/ in "good" or voice consonant pair /th/ as in "they" make sounds using the vibration of the vocal cords. Unvoiced consonants such as /p/ as in "put" or /t/ as in "ten" only use the teeth, tongue, and lips for the pronunciation.

Words that have low energy sounding consonants such as /s/ and /p/ in "stop" or nasal-sounding consonants such as /n/ or /m/ as in "none" or "more" can be troublesome. Like voiced consonants, vowel sounds use the vocal cords. Both produce more energy than unvoiced consonants. Also, be cautious of multi-syllable words that contain weak regions within the word such as "stop system" that contains a /p/ and an /s/ back-to-back in the middle of the word. Some recognizers may view this as silence. If the recognition engine/front-end microphone processing cannot compensate for the environment, task-related vocabulary word phonetic features will make the recognition more challenging.

3. Confusability Performance Testing

Rather than depend on manufacturer performance data, the VCSS project should do an independent test. Typically, recognizer manufacturers will provide performance information about their product. However, most performance data occurs in a benign/controlled environment with trained speakers. The preference is to perform the test on the selected vocabulary in both a benign and a simulated/actual environment. The benign environment test is performed first to baseline recognition performance and to understand the recognition engine's constraints and behavior response to spoken commands. The simulated environment should be a reverb chamber with the ability to pump background noise during the testing. This performance testing is a valuable opportunity to analyze the behavior of the SRT for the given vocabulary words in the given environment-particularly "noisy" words that get recognized easily in the presence of background noise. Again, enlisting the services of the speech recognition engine vendor is highly recommended.

Standard recognizer performance testing does not exist. However, use of a confusability matrix (as shown in Table 2) helps determine which recognizer will provide the performance/features needed for the application. The test provides valuable information on how well the recognizer recognizes spoken in-vocabulary words(words part of the vocabulary) and how well it rejects out-of-vocabulary words (words not part of the vocabulary). Correct recognition responses occur down the diagonal part of the matrix. Substitution errors fall along the off-diagonal elements. Rejections and deletions can also be recorded on the table. The preference is for the vocabulary consist of words in a specific node (a subset of the entire vocabulary) as it is going to be used in an application. The matrix data also serves as reference data in an application to resolve recognition issues such as substitution errors within a state.

If rejection testing is done, the test should document the rejection threshold of the recognizer and the test set used (in and out-of-vocabulary words, coughs, filler words like "um," " ah," etc.). It is highly recommended to repeat the test of the vocabulary enough times to make the data statistically significant. Note that the confusability matrix assumes that each command word is one utterance and not a sentence. Multi-word commands such as "go-to-sleep" or " activate voice loops A and B" are considered one utterance.

Table 1. Vocabulary Confusability Matrix

Predicted Response	Actual Response				
	Word 1	Word 2	Word 3	Word 4	Word 5
Word 1	x				
Word 2		x			
Word 3			x		
Word 4				x	
Word 5					x

Several recognition performance metrics can be obtained from the test. Table 3 shows the scoring types to help in assessing a recognizer. Word Error Rate (WER) and Word Recognition Rate (WRR) are the common metrics used in scoring the performance of the recognizer when used in transcription/dictation since it measures the recognized word sequence against the spoken word sequence. Note that rejection scoring is an assessment of the system for both Inter-vocabulary and out-of-vocabulary word rejections. See rejection errors.

Table 2 Recognition Performance Scoring Calculations

Recognition Performance (%)	Calculation
Recognition Accuracy	$\dfrac{(\# \text{ Correct Words Recognized})}{(\# \text{ Test Words Spoken})} \times 100\%$
Insertion(I)	$\dfrac{(\# \text{ Insertion Words})}{(\# \text{ Test Words Spoken})} \times 100\%$
Substitution(S)	$\dfrac{(\# \text{ Substituted Words})}{(\# \text{ Test Words Spoken})} \times 100\%$
Deletion(D)	$\dfrac{(\# \text{ Deleted Words})}{(\# \text{ Test Words Spoken})} \times 100\%$
Rejection	$\dfrac{(\# \text{ Rejection Responses})}{(\# \text{ Test Words Spoken})} \times 100\%$
Word Error Rate (WER)	$100\% \times \dfrac{(S+D+I)}{(\# \text{ Test Words Spoken})}$
Word Recognition Rate	$1 - WER$
Command Accuracy	$\dfrac{(\# \text{ of commands correctly recognized and processed w/ correct feedback})}{\text{Total number of commands spoken}} \times 100\%$

As a cautionary note, in usability testing, command accuracy can be more important than recognition accuracy. It provides a measure of a performance metric of the system in the actual task environment. Note that both the correct command and correct system response must occur. The command was either successful (100%) or not (0%). Therefore, even though a recognizer scored high in recognition accuracy, it may fail to provide a suitable command accuracy [22].

B. Dialogue Design

The C2 dialogue management gives the VCSS the behavior or "look and feels" of the interface to the user. A C2 dialogue is more user-driven, spending more time issuing commands and focused on completing complex tasks. A major goal of the dialogue is to ensure the task keeps moving forward and errors are mitigated. Even small amounts of errors due to failed recognitions can have a significant impact on time-critical applications [19]. Audio prompts are a critical component of the dialogue [21] as it provides quick feedback means to the user on recognition/task status. Designed correctly, the dialogue will provide a high degree of user acceptability/usability. To meet the high degree of user acceptability/usability requires a synergistic interaction between the user and the VCSS to accomplish the task efficiently and effectively. Situational awareness must also be a design consideration of the VCSS that contributes to what the user needs to perceive and understand as well as helping the user project the future state of the system. [15]

1. Dialogue Structure

The overall C2 dialogue flow logic should be a deterministic state node machine to restrict the command words of the task and reduced recognition errors. The grammar and rules determine the flow logic of the interface. Each state node contains a sub-command word vocabulary relevant to the task for that state node. Transition in and out of each state node occurs through transition voice commands that are located in each state node to permit the user to transition to other state nodes. Within each state node, the user should be able to exit out of the application task and into a standby state node. An example of a dialogue structure is shown in Figure 12.
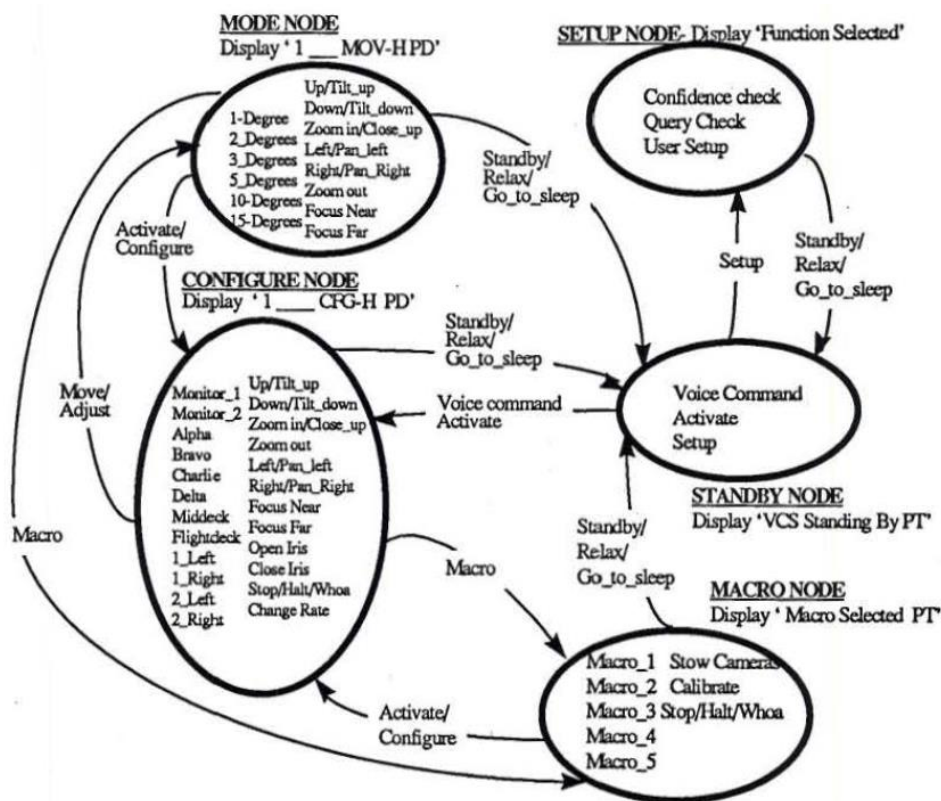

Figure 12 STS-78 Voice Command System Dialogue Structure

C. Activation and Readiness

Background noise and crew/mission control communication chatter could potentially inadvertently activate the system. Therefore, consider using an "arm and fire" command structure to ensure the system task command and control (C2) dialogue is reliably activated when needed. The "arm" command should be a robust word that causes the C2 dialogue to transition to a state node that awaits the "fire" command within a certain time period. If the "fire" command is not annunciated within the specified time, the system returns to listening for the "arm" command. Auditory or visual or both readiness indication should be provided to the user that the system is ready to accept voice commands. In case of background noise inadvertently triggering the recognizer, consider using a push-to-talk electrical switch. In addition, in the event an "arm-and-fire" commanding structure is not working well, consider providing a toggle switch to transition from standby to commanding the system—keeping in mind that the system should return to where it left off and provide feedback to the user where it is at.

D. Deactivation

Consider a means to deactivate the voice control of the task through a meaningful multi-syllable command word, distinct from the other vocabulary words. For gloved-box operations such as dictation/taking notes, consider a VCSS microphone input switch to disable the voice into the recognizer might need to be located inside the glove-box. When the user needs to speak to another crew member, s(he) can hit the switch to disable communications with the recognizer. Like activation, consider providing a toggle switch and associated feedback to transition to a standby/non-commanding mode.

E. Transition Command and Query Words

For C2 applications, being able to move to any of the state nodes is important. Therefore, state node transition command words and associated feedback prompts should be made available at each state node.

Consider an alternate set of transition command words/phrases that perform the same function as the primary command word vocabulary. These alternate command words permit backup commands in case the primary commands are not working well for the user. For example, the command "standby" to place the system in a standby state node can also be accomplished with the command phrase "Go to sleep."

For applications that contain confirmation queries (Ex: Did you ask for system status?) or confirmation of executing a specific command (e.g., the system only expects a "yes" or "no,"), the system should have those command words available at each state node.

F. Guidance on Task Feedback

User task feedback is critical to the success of a speech interface. For space applications, feedback also serves as a safety feature to allow recovery of an unsafe configuration or inadvertent command initiated by the user that could result in confusion if inadvertently taken to another state node. Feedback can be a visual display or an audible tone or speech message. Without feedback, the user may assume the spoken command was recognized and executed correctly. Lack of feedback can cause user frustration and errors. Factors such as consistency and length of the feedback also affect the user [20, 21]. The following guidance for C2 application feedback design should be considered:

- Provide a display to reduce task workload memory. The display provides a visual cue of what was heard, where the user is in the application, and any errors.
- Use short tones to convey quick feedback to the user. Each tone should be different. For example, a correct recognized command and an error tone should be different.
- Speech messages should be kept short and concise. Again, the focus is to keep the task completion, moving forward at an acceptable rate.
- For commands that require more than a few seconds of wait time, such as macro commands, the feedback should provide a periodic indication to the user that the command is in execution. For example, a display would provide a running tally of percent complete of a command.
- A mix of visual display and tones/messages are preferred in a task.
- Provide a mix of a tone and a speech message to aid the user in using the system.
- Since most recognizers need to indicate when they are ready to receive a spoken command (end-point detection), the system should provide a reminder to the user to speak after the tone.

VII. HANDLING RECOGNITION ERRORS

Despite good human factors design and the best SRT, there is a likelihood of recognition errors occurring during the use of the system, particularly in a noisy environment. Errors in an application can affect the user's mental model that contributes to user frustration. This frustration can contribute to error amplification [3] where one error such as a state node error (in the wrong state node issuing commands from another state node) leads to subsequent recognition errors. Therefore, a VCSS should be designed to tolerate the various types of recognition errors that can occur.

Correcting and recovering from an error takes time away from completing the task while specific procedures in the task may not tolerate having to deal with correcting the error. A key aspect of the interface is how the system responds to human error [20]. The speech recognition C2 dialogue should ensure that the user speaks the correct command words in the correct state node. Also, prompts should guide the user when to begin speaking a command word. What follows are the various types of recognition errors that can occur that the designer should consider in a C2 application.

A. Substitution Error

Substitution error is an instance of an error in which one word in the command vocabulary is incorrectly recognized as another command word in the command vocabulary. For example, the user says "status," and the system recognizes "six."

Consider designing system feedback to catch the error and provide options to the user, such as canceling the command. A substitution error cannot be detected by a recognition system. The best line of defense is to consider error reduction by design such as the selection of a small and dissimilar vocabulary derived from the confusability matrix test for the given state node.

B. Rejection Error

Rejection error is an instance of an error due to either correctly spoken command utterance resulting in an inter-word rejection where two words are considered similar, and the recognizer cannot distinguish between the two or correct rejections of out-of-vocabulary (OOV) such as non-vocabulary words, grunts, or coughs. Inter-word is when two words in the vocabulary return similar scorings, and the system cannot clearly distinguish between the two.

For OOV rejections, consider prompting the user again to repeat the command. The assumption is that the system is listening for commands and the user is speaking to the system but may have coughed, or background impulse noise may have been mistaken for a spoken command. For inter-word rejections, the system could return words it thinks it heard (top 2/3 based on confidence scores) and let the user pick to resolve the conflict. If the recognizer has adaption capabilities, the system would update the user template for that word accepted. Rejection errors can occur due to end-pointing parameters set incorrectly, background noise, strong dialect, or speaking to the crew or mission control while the system is listening. For the latter, the preference is to disable the recognizer.

C. Deletion Error

Deletion error is an instance when the system incorrectly omits a component of a valid command string and does not respond. For example, the user says, "Select camera alpha" and the system recognizes only "select camera."

Consider confirming the spoken string via feedback to catch deletion errors. The system could ask for repeat command, or the dialogue provides a different mix of command word/phrases that it will execute on. For example, in the above situation, the system would recognize "select camera" but would ask for which camera since that portion was not recognized.

D. Insertion Error

Insertion error is an instance of an error occurring due to either spurious noise or illegal utterance outside the vocabulary falsely accepted as a legal command word. For example, during high background noise levels, the system captures what it thought was a valid command word. Alternatively, user speech disfluency (any of various breaks, irregularities within the English language, or meaningful non-verbal sound that occurs within the flow of otherwise fluent speech.) results in the recognizer detecting a valid spoken command.

Consider using a recognizer that contains different acoustic models to deal with the noise and provide a means to inhibit the recognizer when the user is not commanding the system. The system should also provide feedback to the user.

E. Spoke-too-Soon Error

Spoke-too-soon error is an instance when the user spoke before the system was ready or background noise was detected as speech. These occurrences usually happen when the user is not notified when to speak and/, or the background noise levels are high.

Consider designing in a short prompt along with a tone or beep sound to indicate to the user the system is ready to accept a spoken command. Alternatively, a push-to-talk(PTT) button could be used, particularly in noisy environments. However, the system should still indicate via visual/tone that it is listening.

VIII. Usability Evaluation and Testing

Despite using the best human-centered design processes, there is still the likelihood that the VCSS design has flaws. Therefore, usability evaluation and tests are a vital part of the human-centered design process and are highly recommended during the development phase of the project [23]. This section provides usability evaluation and test considerations as part of the VCSS development and certification effort. Though the assessment and test can be time-consuming and expensive, the return on value has been shown to be as much as 25-35% improvement [24]. An HSI practitioner should lead to usability evaluation and test development planning.

A. Test Planning and Scheduling

Test planning should begin early in the development lifecycle for both usability evaluation and testing. The development schedule should permit numerous evaluations and test sessions. Also, scheduling should allow enough time between evaluations and test to make changes based on an evaluation or test outcome. Usability evaluations are informal tests to evaluate candidate systems or continue to evolve a selected design after the preliminary design review. After a critical design review, plan on doing several crew evaluations in a high-fidelity environment to further refine the vocabulary and dialogue management. The final tests are usability run-for-the-record testing to verify/validate requirements.

B. Performance Metrics

VCSS performance metrics developed for usability evaluation or testing should be observable and quantifiable. They should disclose the VCSS interface effectiveness (completing a task in a certain time), efficiency (effort required), and satisfaction (How happy was the user with the use of the system?) [22]. In addition, for a VCSS, how many errors that caused a loss of efficiency or task failure should be captured. Note that errors in speech recognition and C2 dialogue prompting/feedback should be uncoupled, i.e., tested independently, as excellent recognition with a poorly designed C2 dialogue could result is out-of-vocabulary words is no better than a poor performing recognizer hiding an effective C2 dialogue design [3].

C. Test Facilities

A usability lab should be available for performing usability evaluation and testing. The lab should be equipped with instruments and data capture devices that the HSI practitioner recommends for evaluating/test the VCSS. Early on, the VCSS is a prototype or engineering model. For usability testing to formally verify and validate requirements, the VCSS should be a flight equivalent unit or flight unit reserved for testing and the environment should be as close to the actual one as possible.

IX. CONCLUSION

The guidelines and considerations presented here provide a project team with a systematic approach to developing a C2 speech application for space. Though speech recognition technology has improved considerably, for space applications, it presents many mission challenges to overcome. However, a reliable and robust C2 VCSS is possible with a human-centered design process, iterative development usability evaluation/test, careful vocabulary selection, and dialogue management design and grammar rules.

Reference

[1] Forberg M., 2003, Why is Speech Recognition Difficult, Chalmers University of Technology

[2] Kuldip P. Kuman , Yao, Kaisheng, Robust Speech Recognition Under Ambient Conditions, Ambient Intelligence, Chapter 6, 2010.

[3} Balentine, B., Morgan M.P, 2007, How to Build a Speech Recognition Application, Enterprise Integration Group, Sam Raven, California,393 p.

[4] Burghduff, R.D. and Lewis, J.R., "Man-machine Interface and Control of the Shuttle Digital Flight System". In Proceedings of the Space Shuttle Technical Conference, Houston TX, 1983.

[5 Griffin, B.N., Hudson P., "Smart Space Suits for Space Exploration," Mars: Past, Present, and Future, Progress in Astronautics and Aeronautics, pp. 297-306, Chapter 27,1991.

[6] Payette, J., " Advanced human-computer interface and voice processing applications in space," in Proceedings of the Workshop on Human Language Technology, Association for Computational Linguistics, Plainsboro, NJ, pp. 416–420,1994.

[7] Yu C., Hansen J. H. L., "A Study of Voice Production Characteristics of Astronaut Speech During Apollo 11 for Speaker Modeling in Space". Journal of the American Statistical Association,2017.

[8] Kim, S., Raj, B., Lane, I, "Environmental Noise Embedding For Robust Speech Recognition," Carnegie Mellon University, 2016.

[9] LaBel A. Kenneth,., Radiation Effects on Electronics 101,https://nepp.nasa.gov/DocUploads/392333B0-7A48-4A04-A3A72B0B1DD73343/Rad_Effects_101_WebEx.pdf,200

[10] Bejczy A. K., Dotson R. S., Brown J. W., Lewis J. L., Voice Control of the Space Shuttle Camera System, 1981.

[11] Author S., Speech Production System, http://www.authot.com/en/2016/08/04/speech-production-system/, 2016.

[12] C. A. Simpson, "Speech variability effects on recognition accuracy associated with concurrent task performance by pilots," Ergonomics, vol. 29, no. 11, pp. 1343–1357, 1986.

[13] Shrawankar U., Thakara V., "Adverse Conditions and ASR Techniques for Robust Speech User Interface," IJCSI International Journal of Computer Science Issues, Vol. 8, Issue 5, No 3, September 2011

[14] T. Yoshioka, A Sehr, M. Delcroix, K. Kinoshita, R. Maas, T. Nakatani, and W. Kellennann, "Making Machines Understand Us in Reverberant Rooms: Robustness against Reverberation for Automatic Speech Recognition, " IEEE Signal Processing Magazine, vol. 29, no. 6, pp.114-126, Nov. 2012

[15] Endsley, M. R.,). "Toward a Theory of Situational Awareness in Dynamic Systems.", Human Factors Journal 37(1), 32-64, 1995.

[16] MIL-STD-1472G Department of Defense Design Criteria Standard Human Engineering (2012, January)

[17] National Aeronautics and Space Administration. (2010). NASA/SP-2010-3407 Human Integration Design Handbook.

[18] ISO 9241-210- Ergonomics of human-system interaction, Part 201- Human-centered design for interactive systems, 3-15-2010

[19] Finan R., Cook M., Sapeluk A., "Speech-Based Dialogue Design for Usability Assessment," IEE Colloquium on Interfaces - The Leading Edge (Digest No.1996/126)

[20] Junqua, J., Robust Speech Recognition in Embedded Systems and PC Applications, Springer International Publishing, Chapter 3, 2000.

[21] Kamm C., Helander M., "Design Issues for Interfaces using Voice Inputs," Handbook of Huma-Computer Interaction, 1997

[22] Pearson, G., "Speech Recognition-Defining System Requirements," Adacel White Paper, 2016.

[23] Tullis, T., Albert, B., Measuring the User Experience, Morgan Kaufmann Publishing, 2008.

[24] Nielsen, J., Iterative user-interface design, Computer, Vol. 26, No. 11, 1993.

[25] Salazar, G., STS-41 Voice Command System flight Experiment Report, https://ntrs.nasa.gov/search.jsp?R=20180001751&hterms=STS-41+voice+command&qs=N%3D0%26Ntk%3DAll%26Ntt%3DSTS-41%2520voice%2520command%26Ntx%3Dmode%2520matchallpartial, 1991.

[26] Salazar, G., Gaston D.M., Haynes D.S., DTO-675: Voice Control of the Closed Circuit Television System, 27]https://ntrs.nasa.gov/search.jsp?R=20180001752&hterms=DTO-675&qs=N%3D0%26Ntk%3DAll%26Ntt%3DDTO-675%26Ntx%3Dmode%2520matchallpartial,1996

[27] Shneiderman B.," The Limits of Speech Recognition," in Communications of the ACM, vol 43, No 9, pp.63-65, September 2000.